DONALD T. HAWKINS
and
B. A. STEVENS
Libraries and Information Systems Center
Bell Telephone Laboratories
Murray Hill, New Jersey
and
A. R. Pierce
Newman Library
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Computer-Aided Information Retrieval In A Large Industrial Library

This paper describes some of the experiences we have had with computer-aided information retrieval during the past years in the Bell Telephone Laboratories library network. In order to put the environment in which we work and the methods we use into perspective, a brief overview of Bell Laboratories and its library network will be given. Following this, some of our methods of information retrieval will be discussed in detail, including machine-readable output and computer-aided literature searching (both batch and on-line). After a short description of our indexing and dissemination methods, we will offer some suggestions on machine searching and draw some general conclusions.

Bell Laboratories is the research and development unit of the Bell System. It fulfills its responsibility for providing new communications systems and services and business information systems by carrying out research and

development and systems engineering. The corporation is jointly owned by the American Telephone and Telegraph Company, which is the parent company, and the Western Electric Company, which is the Bell System's manufacturing organization. Bell Laboratories' annual budget is more than $600 million. It employs more than 16,500 persons stationed at eighteen locations. Many are located at Western Electric manufacturing facilities, where Bell Laboratories works closely with Western Electric in the final development and production of communications equipment.

Of the 16,500 Bell Laboratories employees, more than 7,000 are graduate scientists and engineers. A large number (approximately 2,000) of staff members hold the Ph.D. degree. The technical interests of the staff are very broad, including chemistry, physics, materials science, mathematics, computer and information sciences, psychology, electronics and electrical engineering, speech and acoustics, education and, of course, telecommunications. The company can therefore be viewed as a large and diversified community with a wide spectrum of information needs.

To help meet these information needs, a library network extending to all locations has evolved. The network holdings comprise approximately 150,000 book volumes and over 3,000 journal titles, 2,100 of which are current subscriptions. The network concept is emphasized in all library systems and services. For example, the resources of the twenty-eight units in the network are available to all. Those resources are visible in a printed book catalog which lists the holdings of all libraries in the network. Operating in a network mode means that services can be provided either locally or from a central point to all locations, depending on which method best suits a given situation. Most of the services provided by the library network are the traditional ones familiar to those engaged in library work. This paper will focus mainly on one of these services—the literature search—with a few preliminary words about reference service. (We are primarily concerned here with the outside literature, and not patents, engineering specifications, or other internal documents.)

Reference Service

The reference service is provided by reference librarians at those locations where the demand is heaviest. It is a local service, and as such is close to the user. It is well equipped to provide rapid answers to questions such as: Can you recommend a not-too-technical book on lasers? Which authors have cited my papers in the last two years? What is the current average wholesale price per pound of grey iron castings? Can you provide me with a detailed description of how a sphygmomanometer works? When a question is encountered which requires lengthy searching or specialized technical knowledge, it is passed to the literature searching service.

Literature Searching Service

The literature searching service is staffed by information scientists who compile bibliographies derived from the published scientific, engineering and business literature. The information scientists all have doctoral degrees in a technical field. Searches are undertaken in any area of need. The literature searching service is centralized at one location, and an attempt is made to have as many pertinent abstract journals as possible available there. An example of the diversity of search requests is shown by the titles of the following recent searches performed: aluminum joining, multidimensional scaling, mechanical properties of gold, infrared testing of integrated circuits, light scattering from fluids, minicomputer software, radiation effects on polymers. Figure 1 shows the subject distribution of ninety-three searches conpleted in an eight-month period.

Search Methods

This brings us to a description of the methods used by the information scientists in answering search requests. The possibilities are: (1) searching published abstracts and indexes manually; (2) using commercial searching services in the batch mode, receiving results in hard copy or machine-readable output; and (3) searching commercial data bases through an on-line terminal. Until 1972, traditional manual methods were used almost exclusively. Since then, however, a basic change has taken place, and machine methods have grown rapidly and now are the most often used. Hawkins's survey of the data bases used by the information scientists which covered the period from January 1972 to May 1973 found that machine retrieval methods were used in about one-third of the searches;[1] the figure is now well over one-half and is growing.

Machine Searching

Machine searching is the subject of an increasing number of publications. We will summarize what appear to be its major advantages and disadvantages. The advantages are:

1. Machine searching is exhaustive, but not exhausting. A machine can "see" all instances of a term in a data base. It can rapidly scan a large volume of information, and it does not get tired.
2. Normally, computers operate rapidly. We say "normally" because (on-line) system response can be degraded if a large number of users are trying simultaneously to gain access to the same data base or system, or if one user is tying up major resources. One must also be aware of

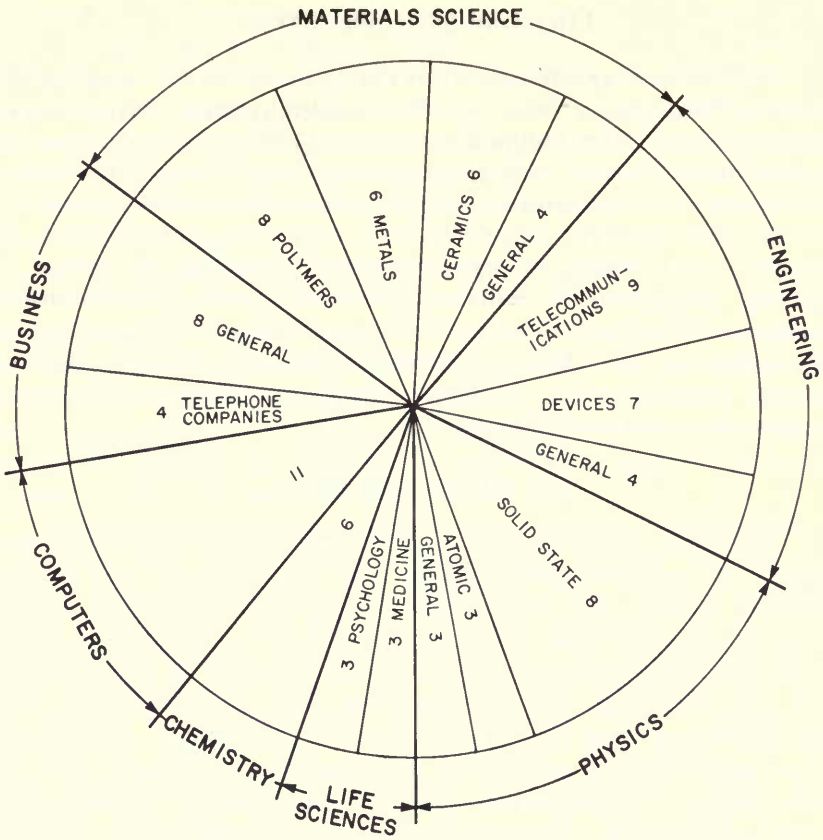

Figure 1. Subject Distribution of 93 Literature Searches Conducted Between April and November, 1973. (Figures indicate numbers of searches.)

system malfunctions, restarts, etc., which make the machine unavailable for periods of time varying from minutes to days.

3. The output is legible and well formatted, so that normally we do not have to expend clerical effort to edit or repackage it; the information can be given directly to the user. We frequently satisfy information requests with a list of references printed at the terminal and given directly to the requester.
4. There is the possibility of obtaining machine-readable output for further processing.

One disadvantage of machine searching may be the purchasing of services from outside suppliers. However, these disadvantages are minor annoyances

accompanying a very advantageous facility. The negative aspects of obtaining machine searches from outside suppliers arise from the immaturity of the industry: suppliers are volatile. They tend to set their rates low in the hope of attracting a high volume of business, and then abandon the business when the volume does not meet their expectations. After deciding that a supplier offers a good service, investing time to study the system, and possibly writing interface programs to make the output compatible with your system, you may find after some months that the tape format has changed, or discover that the supplier cannot be depended on to make deliveries, or that they have simply shut up shop, and the investment made in adapting to the offering is wasted.

Assuming that a supplier is viable, we can list some of the other disadvantages of machine searching.

1. Not all important data sources are available. This disadvantage is being alleviated by the rapid growth in the number of computer-readable files being offered to users.
2. The data cover a limited time span. Very few files have information preceding the late 1960s: the government reports (National Technical Information Service) and American Petroleum Institute's files are notable exceptions. It might be expected that some data base producers would extend their files back in time, which would greatly aid retrospective searching. However, the clerical effort needed merely to keep files current is enormous and, with rising labor costs, it would appear that there is very little likelihood of such a backward extension. Cuadra observes from the standpoint of a search vendor that older data may not generate enough usage to justify economically keeping it on-line.[2] Some on-line systems drop off older data as new data are received, keeping only a fixed amount on-line. This may be useful for a market information file, or for a file listing new developments in a fast-moving technical area, but it severely limits its usefulness for retrospective searching.
3. The relevance ratio is sensitive to the search strategy, particularly in a batch system where one is unable to review the results as the search proceeds. Computers can generate a large volume of paper in a very short time, and one must use care in defining a search strategy which makes the output meet the requester's needs.
4. Finally, costs can be high, especially when many items are retrieved and one must pay a per-item charge. However, the costs of machine searching are often much lower than the costs involved in manual searching. Normally, therefore, cost is not a negative factor. Elman has recently studied this point, using the DIALOG system, and comes to the conclusion that the cost of the "average" manual search (if such a thing

exists) is $250, compared with $47 for the same search done by machine.[3]

We should point out that machine searching is not a cure-all. It cannot be used in every case. For instance, where a broad term exists in the thesaurus without a desired modifying or qualifying term, it will be necessary to use the manual approach in which the abstracts or printed index are scanned, rather than just titles. One recent example we encountered was in the preparation of a large bibliography on water,[4] where it was found impossible to separate by machine methods references on water of hydration, natural waters, etc., from those dealing with water as a pure substance or chemical entity.

One of the more hotly debated subjects in the area of machine searching is the question of the need for an intermediary, such as an information scientist, between the original requester and the computer terminal.[5] Our experience has been that such a person is needed to keep up with software changes, new data bases which become available, and changes in existing data bases. Efficient use of the system requires that the searcher know how to search the data, what elements in it are searchable, what output formats are available, etc. He or she must know which data bases allow free text searching, and which use a controlled vocabulary or thesaurus. A scientist or engineer will not normally use a system often enough to develop or maintain the skills necessary to exploit it efficiently. We have found that many original requesters are interested in machine searching, especially on-line searching, and that it is often useful, but not necessary, for them to be present when the search is performed. However, few of them wish to become burdened with details such as those we have just mentioned. We have observed that, as with any new method, users are at first enthusiastic, interested, and even fascinated, but when the learning of details becomes necessary or problems arise, the enthusiasm soon wanes and the user is quite glad to have the information scientist do the job.

We now turn to a description of some of the methods used in our information retrieval activities, focusing on computer-aided methods. Computer methods traditionally are divided between batch and on-line methods. In our own environment, we further characterize batch methods according to whether machine-readable output is obtained. After a general description of machine-readable output, we will discuss each type of searching in detail.

Machine-Readable Output

Increasingly, we find it appropriate to get the output of a search in machine-readable form (e.g., on magnetic tape) as well as, or instead of, in

hard copy. One might ask why we desire machine-readable output. The answer lies in the desire to amortize the high cost of preparing a large bibliography by making the information it contains available to more than one user. Because of the breadth of Bell Laboratories' interests, there is usually a sizable potential audience for the bibliographies we produce. Quite often, a search request which is narrow in scope can be broadened to produce a bibliography of interest to a large number of people. Having search results in machine-readable form enables us to repackage the information into a uniform style, perhaps to add information from other sources, and then to provide an index to it using our permuted indexing system, BELDEX.[6] We also have a few ongoing bibliographies which appear at regular intervals, covering broad subjects of great interest to Bell Laboratories, e.g., optics and circuit theory. Because of their size, preparation of these bibliographies depends heavily on machine methods.

Some of the advantages and disadvantages of using the machine-readable output from a literature search should be identified. The advantages include:

1. It avoids redundant keypunching of data and the possible introduction of errors. Data once entered into machine-readable form do not have to be entered a second time, which avoids duplication of effort.
2. Large volumes of data can be easily manipulated by machine, as we have already discussed.
3. The cost of producing a bibliography is greatly lowered. Apart from the information scientist's salary, the greatest cost of bibliography production is data entry, such as keypunching. Using or modifying existing machine-readable data significantly reduces input costs.

The general disadvantages of machine searching listed above apply to machine-readable output as well. Some other disadvantages peculiar to machine-readable output are:

1. There are problems associated with magnetic tapes. The transfer of information from one computer installation to another is fraught with difficulties. Nothing is standardized—neither tape writing densities, record formats, character sets, nor terminology.

   The supplier must state the characteristics of the tape. Sometimes incorrect information is given, or words are used in a sense that is different from one's understanding of them. One must then ask a computer specialist to read the tape on the local machine. At the worst, he may be unable to do so if, for example, he does not have a tape drive for the appropriate density. Once the tape is mounted on a suitable drive, further traps await. Labeling, logical and physical records, record format (fixed or variable

length, or spanned) and character set all must be properly characterized and handled by a local computer program. Moreover, a tape from an outside installation is likely to require little-used options in local utility programs. These are most likely to contain software errors. Finally, the tape may be physically defective: there may be spots on it where the magnetic coating is defective, or where the tape has been stretched.

If difficulties arise, the local expert may not be able to determine which of the various factors is responsible. Several weeks of interaction between you, the supplier, and the local computer center may be needed to clarify this.

2. Once the tape has been successfully read by the computer, the data must be manipulated by a program to the desired form. You may find that the layout chosen by the supplier for the information fields is difficult to manipulate by your programs. For example, the author field may not distinguish between first names and family names, or between personal names and affiliations. Suppliers are also apt to change their formats or computers, both of which may be troublesome. If the format of the data has changed since the last running of the conversion program, the program will not run and must be altered, causing more delays. Sometimes the changes are not announced by the supplier, so that a debugging process must first occur. Errors in the data often occur and must be corrected.
3. Still another disadvantage of machine-readable output is the one inherent in the loss of control over any process which depends on an outside supplier. If the supplier is having trouble with his tapes, his service bureau, or the data base producer, search results can be considerably delayed.

In spite of the disadvantages associated with using machine-readable output, we have found the practice to be most useful. We reiterate our opinion that the advantages far outweigh the disadvantages, especially after the initial difficulties have been surmounted and the process is on a production basis.

Interactive Retrieval

We now turn to a discussion of some of our experiences with interactive information retrieval. Our experience has been limited to the retrieval systems of Lockheed Corporation's DIALOG and System Development Corporation's ORBIT. We are relative newcomers to the ORBIT system, so that the following discussion mainly concerns the DIALOG system's behavior under our probing of its files. Our particular interest in the DIALOG system is based

	Average Values	
	Per Session	*Per Search*
Duration	21.2 min.	44.6 min.
Cost	$21.10	$44.50

Table 1. DIALOG Search Statistics, March-December 1974
438 Sessions, 208 Searches, 2.1 Sessions/Search

largely on its coverage of electronics, physics and computer literature through the INSPEC (*Science Abstracts*) data base.

We will now explore in detail some of the characteristics of our search experience. Table 1 gives a thumbnail sketch of the time and money spent during a ten-month period on the DIALOG system. We recorded 438 interactions with DIALOG on 208 search topics for 2.1 sessions per search. This is because many searches require data from more than one data base, and also because after reviewing the results one frequently gets a new idea and finds it profitable to return and try a new tack. These sessions averaged about twenty-one minutes, so that a search itself averaged forty-five minutes. This figure is in exact agreement with Elman's results. Further, after costs for TYMSHARE, batch print-out, and connect time are computed, we find that the rule of thumb "a dollar a minute" is a remarkably accurate gauge of how much one is going to spend on a search.

Of course, "average" figures may not be "typical." Figure 2 is a histogram of session durations for the 438 sessions included in this sample, and shows that the most frequent duration interval is somewhere between five and ten minutes and that, overall, 60 percent of the sessions took less than twenty minutes. It is those few very long searches that boost the average to twenty-one minutes. Total cost (see Figure 3) looks quite similar to the session duration histogram. Again, a few expensive searches—or a few generating voluminous output—boost the average cost.

We now turn our attention to what we were looking for in these 438 sessions. Figure 4 shows how the sessions were divided with respect to search objectives. The term *objectives* seems appropriate here because many of the sessions were concerned with objectives other than the traditional one of finding what has been written on a given topic. Notice, for instance, that slightly more than 15 percent of the 438 sessions were devoted to bibliometric research, and to demonstrations of the search system to members of the library staff. We feel that these in-house demonstrations are valuable in alerting our colleagues to the kinds of things that can be done easily using computer-aided searching. Often the knowledge that the facility is there and is

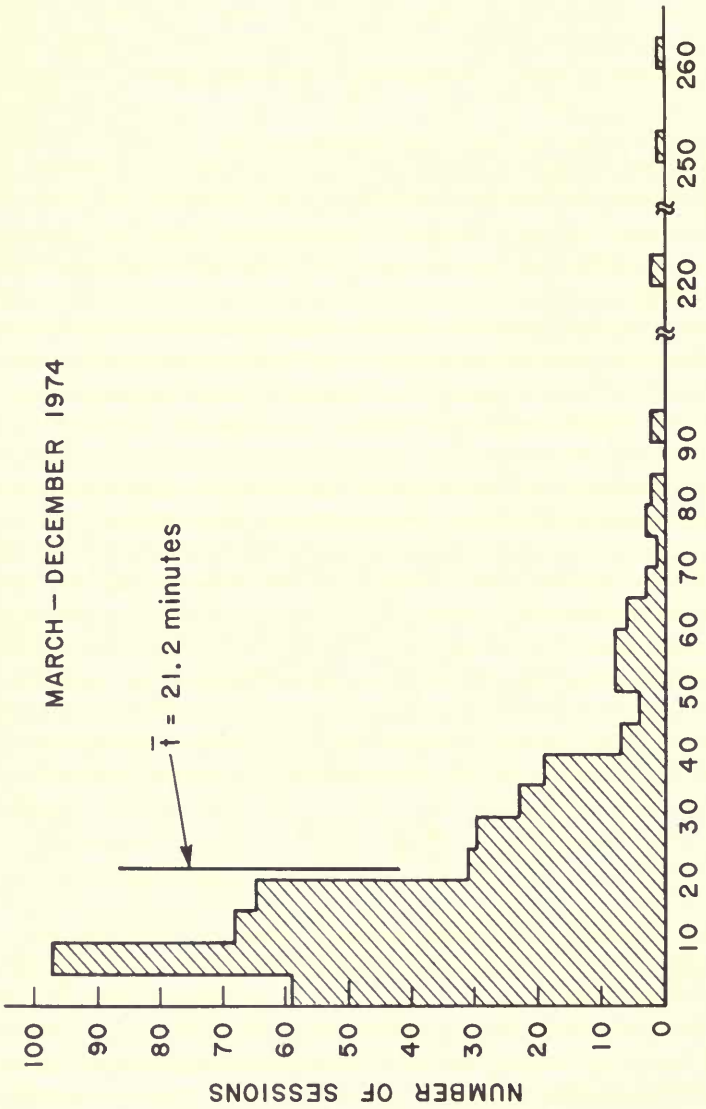

Figure 2. Session Duration Using the DIALOG System—438 Sessions

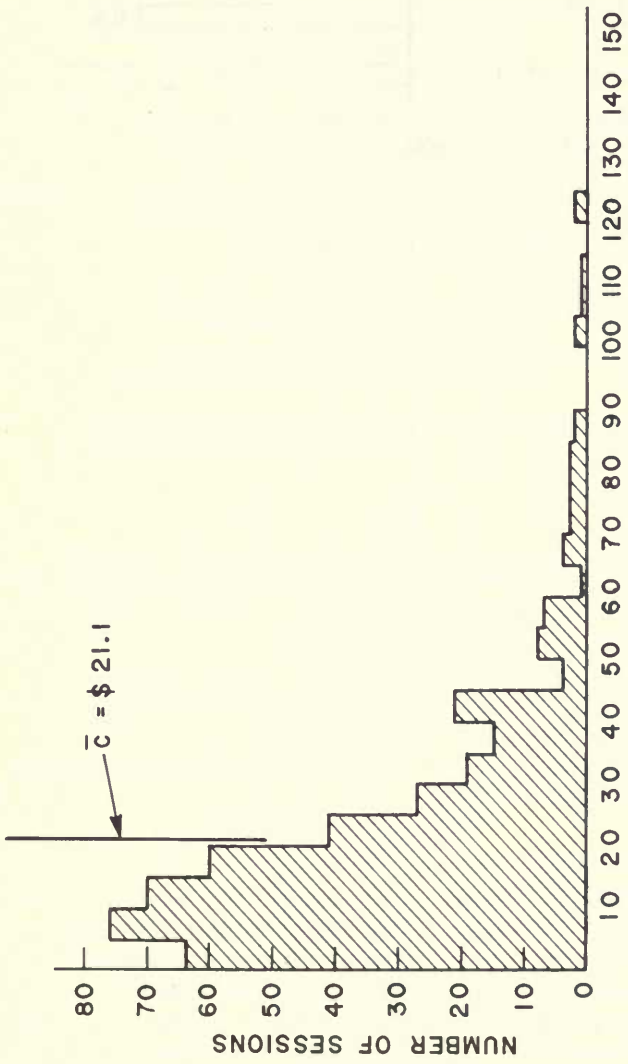

Figure 3. Session Cost Using the DIALOG System—438 Sessions

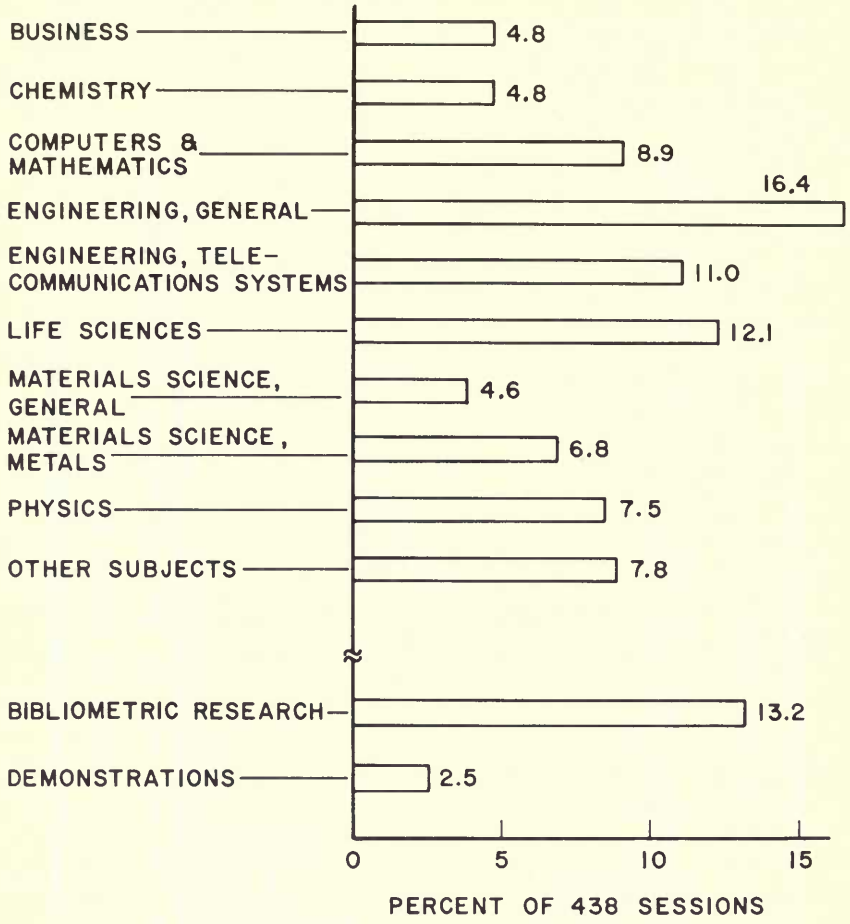

Figure 4. Search Objectives

relatively painless to use will prompt requests that might otherwise be done manually, or not at all. Using an interactive search system as an aid to bibliometric research also makes sense. Consider the following problem: we want to know how heavily various institutions publish in specific areas of science and technology. Using interactive retrieval, it is fairly straightforward to determine how many articles authored by institutions A, B and C, and appearing in core journals X, Y, and Z, were indexed by *Physics Abstracts*. Assuming that what is indexed by *Physics Abstracts* is, at least, an unbiased cross-section of current physics literature, it is then possible to gauge the relative publication activities of institutions A, B and C. Of course, in practice,

there are more than three core journals in any field, and great caution must be used in drawing conclusions: if the number of articles written by A is twice that of B, one certainly cannot say that organization A is twice as "good" or twice as "productive" as organization B. Interesting trends can be found, however, and Bell Laboratories' management has found answers to some of their questions through such bibliometric analyses.

Let us now look at the subject-oriented sessions. The subject breakdown presented in Figure 5 is peculiar to Bell Laboratories' interests; not many other institutions would divide their interests this way. One thing to note is the wide spectrum of topics, few of which, it seems, are directly concerned with anything that is much like a telephone system. There is, of course, a lot more to any research and development effort than the end product, and this accounts for the variety of topics covered. It is interesting to note that while our information scientists have backgrounds in the physical sciences, they cannot control the information needs of their community, so sometimes one has the situation of physicists and metallurgists searching the business and psychology files looking for references on job enrichment.

Figure 5 shows how frequently the data bases that are accessible through the DIALOG system were used for each of the subject areas previously defined. The numbers given for particular subject areas are the percentages of the total number of *sessions* that a given data base was searched. (Because there were 438 sessions in all, 0.2 percent represents one session.) Note that there are peaks along the vertical and horizontal directions. The horizontal peaking is simply a statement that several data bases may cover a given topic, while vertical peaking means that a given data base covers, from our point of view, several subjects simultaneously. From our perspective, for example, *Chemical Abstracts* is more related to materials science than to chemistry. The fact that most data bases exhibit quite a few peaks indicates why it is profitable to search a topic against multiple data bases. Hawkins found that 55 percent of our searches used two or more data bases. Figure 6 shows our total data base usage in the 438 sessions.

At this point, it may be useful to consider some of the problems arising in the retrieval of a document reference using the surrogate appearing in a bibliographic data base. The whole process of indexing and abstracting—by any system whatever—is based on the wildly ambitious assumption that a few keywords, a bit of subject code, and an abstract can represent the information contained in the document. We all know the deficiencies of indexing, cataloging and abstracting, but perhaps not enough has been said about the shortcomings of this particular approach. How much important content is lost by transforming a document into a collection of entry points in some bibliographic data base remains an unanswered question.

Another consideration to be kept in mind is when the point of diminish-

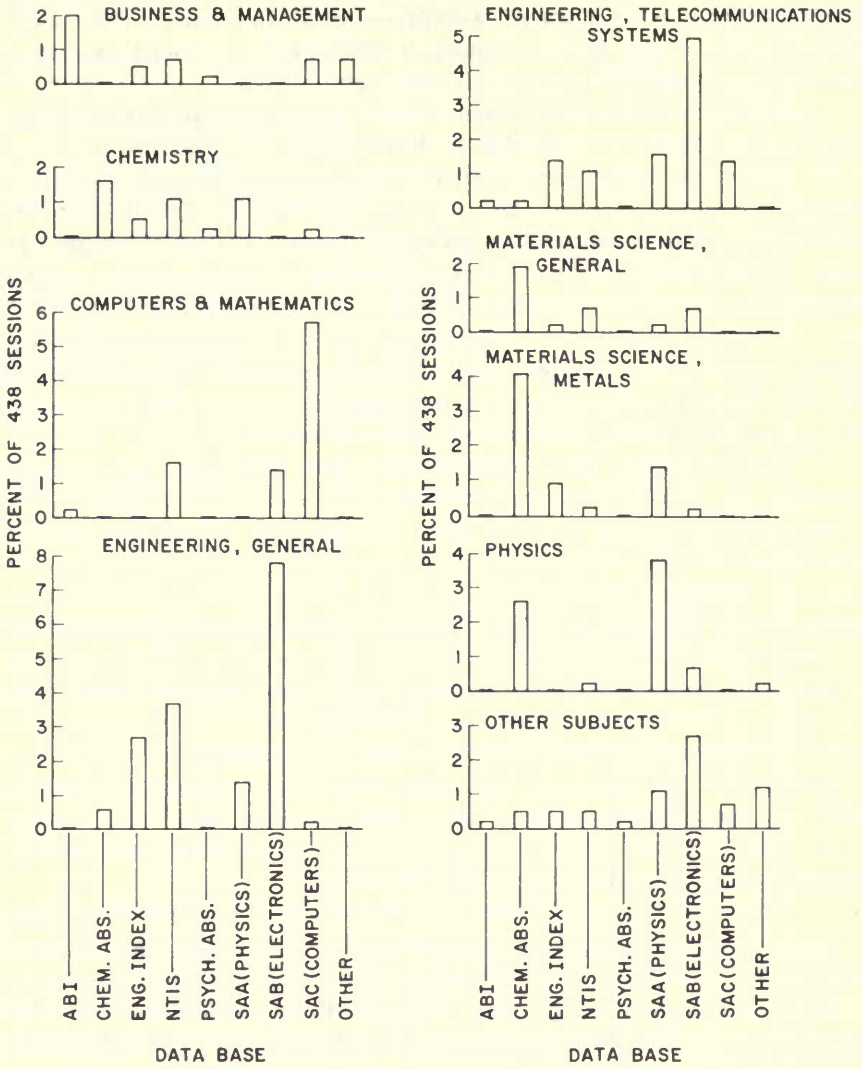

Figure 5. Subject and Data Base Breakdown of 438 Sessions Using the DIALOG System.

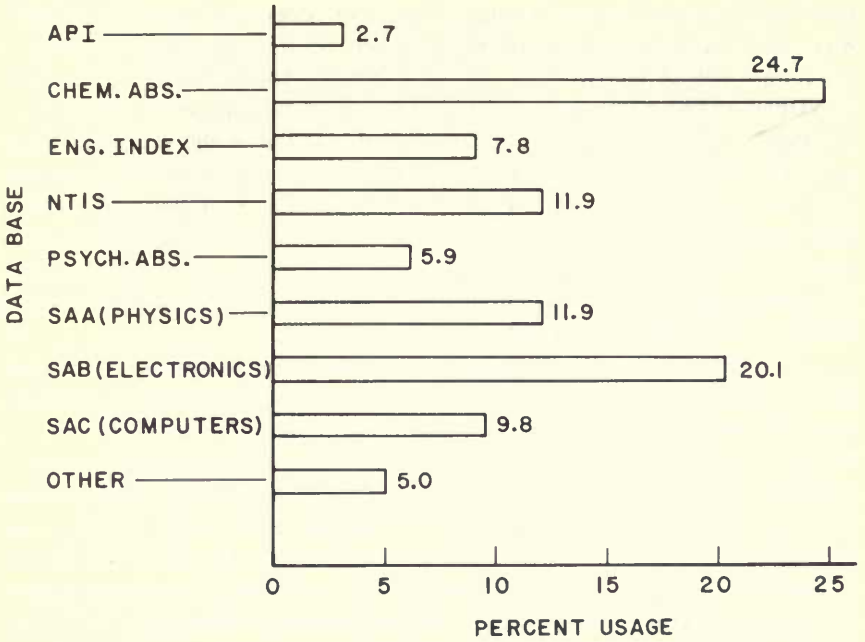

Figure 6. Data Base Usage (438 Sessions).

ing returns is reached; that is: When should a search be terminated? We have already noted that it is often profitable to search more than one data base, but we also know that searching too many data bases results in at best only a few additional references and, more often, considerable additional labor.

When machine search methods are being used, the problem becomes more difficult because of the volume of data that a search may yield. With manual searching of printed sources, items are acquired one at a time and filed; thus it is immediately known if an item is a duplicate. However, a machine search might yield hundreds of items at one time. An example from our recent experience will make the problem clear. Our search topic was from the field of materials science: "soft modes of lattice vibrations in solids." The first choice of data base was *Science Abstracts A*; a search yielded 200 references. A second choice was CA CONDENSATES, which yielded 100 references. By visual comparison of the items, it was found that 34 of the 100 items from CA CONDENSATES had not been retrieved from *Science Abstracts*.

This illustrates the nature of the problem; if the information scientist does not search the second data base, 15 percent of the potential material will be missed. If the second data base is searched, a laborious comparison to

eliminate duplicates must be undertaken. Two questions immediately arise: Why were the thirty-four items not found in *Science Abstracts*? and Why is the comparison of the two sets (one from *Science Abstracts* and one from CA CONDENSATES) a laborious process? The reason for failure to retrieve items is complex; some causes are: (1) the original creation of the data base was at fault; (2) the item is of a type which does not receive exhaustive coverage from the data base supplier (reports, theses, patents, etc.); and, (3) retrieval failed because the item did not have appropriate descriptors. In the last case, inspection of the two output sets may suggest improvements in search strategy but, again, delay and expense are incurred if a new search is undertaken and a further hunt for duplicates must be made.

To understand why elimination of duplicates is expensive in time and effort, we must consider how the items are inspected visually. As they are originally acquired, the two sets of items from the two data bases are not arranged in a sequence that makes an immediate scan possible. Instead, a subsidiary file must be made from one or both, arranged by data elements that are considered least redundant. In practice, this might be the page number of the document. If a match is found on this element, a comparison is made on another one, probably journal title, and finally on volume number. Creating this inverted file for a set of several hundred documents is not cheap, whether machine or human methods are used.

The problems of information retrieval are compounded by the many different kinds of approximation any one data base makes to represent the documents it seeks to announce to potential users. If one tries to search *Chemical Abstracts* on-line he is not searching anything that looks like *Chemical Abstracts* at all. *Chemical Abstracts* on-line is the CA CONDENSATES file, which has far fewer entry points than the *Chemical Abstracts* volume indexes.

The CASIA (Chemical Abstracts Subject Index Alerts) data base does list all *Chemical Abstract* index entry points. However, CASIA has no abstract information. Figure 7 compares the two forms for the same document as they appear in CA CONDENSATES and the printed *Chemical Abstracts*. To further compound difficulties, different search systems usually treat any given data base differently. Although two competing data bases may have many journals and other sources in common, there is the problem that each will represent a given entry differently. Thus a strategy that worked well in one data base may prove to be fruitless when applied to another. On the other hand, the same strategy applied against multiple data bases (if this is possible) may be insurance that less is missed. Because of such hazards, interactive searching is very attractive since one can change logic or data bases as the search progresses, based on results accrued to that point. It is this feature—the ability to learn from mistakes and then take corrective action before the search logic freezes—that makes interactive searching so powerful and attractive a tool.

134257z Improving the color of Ziegler olefin polymers. Ainsworth, Oliver C., Jr.; Lochary, Joseph F.; Stain, Shelton D., Jr. (Dow Chemical Co.) **U.S. 3,773,743** (Cl. 260-94.9F; C 08*f*), 20 Nov 1973, Appl. 796,153, 03 Feb 1969; 5 pp. Olefin polymers contg. ≤500 ppm metallic catalyst residues which characteristically discolor on exposure to high temps. were stabilized against discoloration and degrdn. during and after high temp. processing by intimately contacting the polymer with about 0.5–1.5 wt. % (on polymer) OH compd. contg. 0–12 C atoms and ~50–2500 ppm $C_{\leq 2}$ Lewis base boiling ≥100°, and processing the polymer contg. the alc. and base at a temp. above the polymer softening point to improve its color. Thus, *polyethylene* **[9002-88-4]** prepd. by low pressure polymn. in hexane in the presence of a 1:1 *titanium trichloride* **[7705-07-9]**= *-triisobutylaluminum* **[100-99-2]** catalyst was steam distd. to remove the hexane and a substantial portion of the catalyst, and the polymer contg. 30–50% water was dried to <0.1% moisture to give samples contg. 76 ppm Ti residues, and having Milner color 72.6. Each sample was charged to a feed section of an extruder operating at 190–250°, and as the sample passed into the extruder from the feed section, a mixt. of water and an org. base contacted the polymer. Thus, 200 ppm *calcium stearate* **[1592-23-0]** and 1.4 wt. % water were added to polymer in the feed section, and extruded to give a product with Milner color 86.5. When the Ca compd. was increased to 2000 ppm, the color was 89.6.

```
CA080241342572
  Improving the color of Ziegler olefin polymers
  AUTHOR: Ainsworth, Oliver C., Jr., Lochary, Joseph F., Stain,
Shelton D., Jr.
  SECTION: CA036006     PUBL.-CLASS: P     COVERAGE: 1
  JOURNAL: U.S.     CODEN: USXXAM     PUBL: 731120     PAGES: 5 pp.
  DESCRIPTORS: Lewis base stabilization polyolefin, color stability
polyolefin, degrdn resistance polyolefin, polyethylene color stability
  PATENT-NO: 3773743     APPLIC-NO: 796,153     DATE: 690203     CLASS:
260-94.9F, C 08f
  ASSIGNEE: Dow Chemical Co.
```

Figure 7. Chemical Abstracts and CA CONDENSATES Representation of the Same Abstract. (Reproduced by permission of Chemical Abstracts Service.)

Batch Searching

We use batch searching when the data we require is not available to us through interactive search techniques. For example, not being members of the American Petroleum Institute (API), we search this data base by going directly to API with our problem, and let them do the probing. In one way, not having access to a rarely used file (in our case, the API files are rarely used) is something of a benefit. It is unlikely that we would ever be as effective information intermediaries as the specialist on that file for what is, to us, an exotic file. This example can be generalized to libraries or information centers not having subject specialists on staff. In this case the batch information centers offer the advantage of the review of profiles by specialists.[7]

BATCH RETRIEVAL WITH MACHINE-READABLE OUTPUT

As mentioned before, obtaining search results in machine-readable form results in great economies in the preparation of our bibliographies. Beyond saving the direct costs of keypunching, recall the additional considerations favoring this approach: (1) errors are not introduced through the transcription process, and (2) no one should have to keyboard material that someone else has already keyboarded. The first point is self-evident, and the second focuses on what Weiner called the "human use of human beings."[8]

Figure 8 shows in some detail the steps involved in utilizing machine-readable output. First, the results are received on magnetic tape from one or more sources and converted to a format compatible with BELDEX, our KWIC (Key Word In Context) indexing system, using a FORTRAN program. Next, a BELDEX run is made and the results are reviewed by the information scientist. (More details on BELDEX operations are discussed below.) Typically, a copy of the preliminary BELDEX results is forwarded to the original requester at this point, giving him something with which to get started. Now the information scientist has an opportunity to review the results, delete the irrelevant entries and perhaps to classify the bibliography into sections. Additional data—possibly from either another batch search or manual searching—can be integrated with the original results. For this purpose we utilize a very powerful on-line tool called QED (Quick EDitor).[9]

QED is a string-oriented, interactive programming language. For example, it is easy to find and correct spelling variants using QED. Consider the Americanization of the word *colour*. We require the program to find each occurrence of *colour* and substitute *color* for it. The code to achieve this looks strange but is brief: 1,$s/colour/color/. It is read: "On every line of the file containing the string 'colour,' substitute the string 'color' and now all lines have been Americanized." Such fixes were used in a recent bibliography on color television, for example.

Because QED is a programming language, we can run stored QED programs against new input data that are in nonstandard formats and convert the data to our standard BELDEX form algorithmically.

The interactive feature of QED means that we always know the current status of the file we are working on. Gaps in time between editing and running the edit program do not exist; thus, we are not required to dredge up from a hazy memory a recollection of what was going on. Interactive editing lets us focus our attention on the job at hand. For example, consider the chore of adding subject class codes to items in a bibliography. Frequently, this is done after the information scientist has reviewed the preliminary BELDEX output. This would be a difficult task using punched cards since each card would have to be found and removed, the subject code punched, and the card refiled. Using QED, the editor works through the file using a stored program

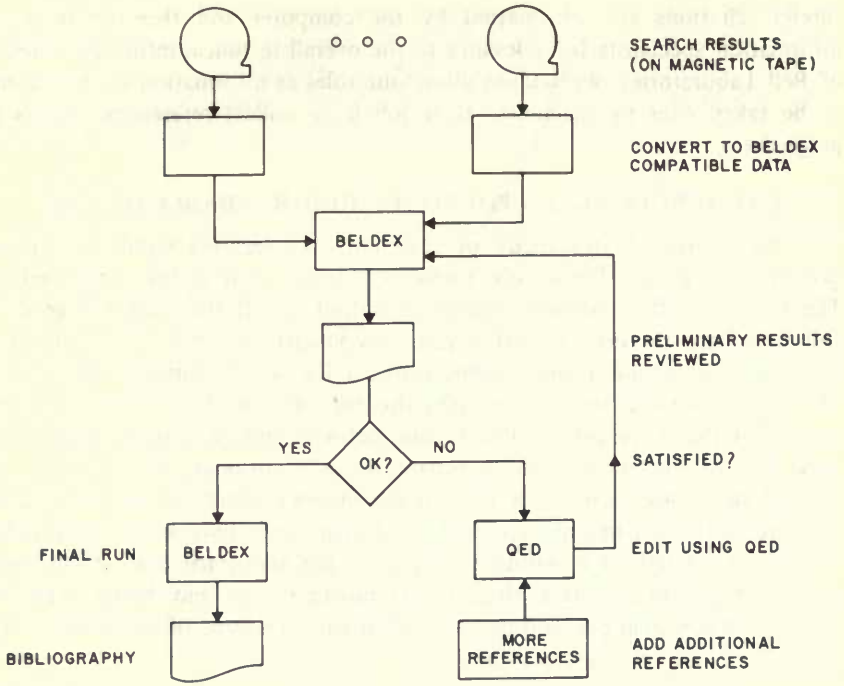

Figure 8. Steps in Bibliography Production.

that requires only enough keystrokes to specify the subject code of a particular item. Occasionally the data the program sees are anomalous, and then it pauses in midstream. For batch processing this would be a disaster and the job would abort, but in time-sharing there is the opportunity to back up, fix the unexpected data string, and continue to a successful conclusion. The process of review and correction continues until the information scientist is satisfied that the product is of acceptable quality.

We also make considerable use of tapes purchased from commercial sources—in particular, the INSPEC (Information Service in Physics, Electrotechnology and Control) tapes. Here we are doing much less ambitious searching than Lockheed, System Development Corporation, the Illinois Institute of Technology Research Institute, and the Knowledge Availability Systems Center offer. Information systems developed at Bell Laboratories for this purpose select all entries from a data base by the journal in which the article appeared. This journal/subject filter scheme is a major part of the input for our largest current awareness bulletin, *Current Technical Papers*. This bulletin appears twice each month in five sections. On the average, the total number of papers announced in all sections exceeds 2,000 per issue. These

filtered citations are reformatted by the computer and then reviewed by information specialists for relevance to the overall technical information needs of Bell Laboratories. We seldom allow our roles as information intermediaries to be taken over by machines: their job is to collect references, ours is to judge them.

BATCH RETRIEVAL WITH NO MACHINE-READABLE OUTPUT

Our usage of this mode of searching has declined rapidly as on-line systems have grown. We do not foresee much use of it in the future, either. The reason for this concerns volume of output: (1) if the output is great, it will be costly to process further (i.e., keypunch), and (2) if the output is small, we may spend a considerable sum for the search, but get little return. This does not mean that we consider this way of operation to be a bad one, but rather that it is not adapted to our methods. Indeed, a batch search with only printed output is quite adequate for a traditional SDI (selective dissemination of information) service, or for libraries which do not have access to a time-sharing computer terminal or system. Such libraries or information centers are spared what would be for them the additional cost of cumputer terminal rental or purchase. High quality profile review, and the fact that one needs no more than pen and paper, explain why this type of batch retrieval is the right choice for many searches.

Dissemination

We now address the question of what to do with a mass of references collected on a given topic in response to a search request. This is not a trivial matter: in many cases we accumulate 1,000 to 2,000 references for a given search.

When the data collection phase is complete, the resulting references may be an unorganized mass, derived perhaps from different sources, with duplicates, inconsistent styles of citation, typographical errors, etc. Some may be recorded on computer printout, some may exist in machine-readable form and some will be on hand-written records. It is useless to distribute this confused mess. Our task is to organize it, index it, and provide easily readable copy in a form that lends itself to reproduction. Our product will then not only serve the original requester optimally, but will also be available to others who might find it useful.

The task is composed of two operations: (1) mechanical–get the words onto the page; and (2) editorial–correct errors and provide an index. As previously noted, we have at our disposal a computer aid, BELDEX, developed specifically for these purposes. It creates good copy with numerous indexes to facilitate editing and utilization as a bibliography.

The input to BELDEX is a machine-readable file of records, each tagged to show its nature, e.g., title, author, or bibliographic citation. Furthermore, the group of records constituting one item of the bibliography must be together, with the last one marked to show that it is the last. BELDEX, when presented with a file of such records, performs the following tasks:

1. Report generation–the creation of the bibliography itself, a complete listing of all the references. The user has great freedom in specifying format.
2. Indexing–BELDEX can create indexes for any data elements. The most important is a permuted title index (KWIC); another very important one is an author index. These two usually form part of the completed bibliography; other optional indexes, such as a source-journal index, are sometimes used as editorial aids.

In many cases, we wish to list the bibliography items in some order other than the order of acquisition or input. Possible choices are by a subject breakdown, by author, or chronologically. BELDEX will create this order by sorting the input material using tags that are present on the items.

However, we feel that one of the more important accesses to the information in the bibliography is the permuted title index (KWIC). We have devoted much effort to developing BELDEX indexing into a powerful tool with the well-known advantages of KWIC while avoiding some of the disadvantages. For example, BELDEX provides a stop-list of nonentry words, such as *a, the,* and *of*; all KWIC programs have such a list. However, in BELDEX, the list is generalized to an "action" list. The most common action is indeed "stop"–do not use the word to create an index entry–but the user also has other options, such as: "go"–create an entry only if the word is on the list; create *see also* references; replace unauthorized terms with authorized ones; or ignore prefixes for indexing purposes. These are some of the built-in options. For special needs, others may be implemented within the framework of the action list.

KWIC indexes are useful for error correction because a typographical error will often be displayed in the index in the context of correct words from other entries. Figure 9 shows how this can be used to detect spelling errors, or inconsistent volume/year information in a citation.

BELDEX was created by the Libraries and Information Systems Center at Bell Laboratories; our center maintains it and uses it daily. Consequently, it is continually evolving. As new needs are perceived, they are incorporated into the system. Over a fifteen-year period, this system has become a flexible, powerful and reliable tool.

```
  LIGHTNING ARRESTERS OF LIGHT WEIGHT CONSTRUCTION.
COAXIAL CABLES STRUCK BY LIGHTENING.
COAXIAL CABLES STRUCK BY LIGHTENING.
HOTOELECTRIC DETECTOR OF LIGHTING.
STERS WITH SWITCHING AND LIGHTING OVERVOLTAGES UNDER NE
UDY OF THE PARAMETERS OF LIGHTNING.
VOLTAGE STATIONS AGAINST LIGHTNING.
                         LIGHTNING AND SURGE PROTECTION
                         LIGHTNING ARRESTER CURRENTS.
TCHES AND PERFORMANCE OF LIGHTNING ARRESTERS.      SWIT
E DESIGN OF HIGH VOLTAGE LIGHTNING ARRESTERS.
CS OF 8.4-KV, VALVE-TYPE LIGHTNING ARRESTERS AGAINST ST
       REPRESENTATION OF LIGHTNING ARRESTERS IN MODEL T
                         LIGHTNING ARRESTERS OF LIGHT W
                         LIGHTNING PERFORMANCE OF OVERH
    THE PROTECTION FROM LIGHTNING STRIKES OF COMMUNICA
    THE PROTECTION FROM LIGHTNING STRIKES OF COMMUNICA
NSMISSION SYSTEMS DUE TO LIGHTNING STROKE ON OVERHEAD G
 KV LINES AGAINST DIRECT LIGHTNING STROKES.
                    HOW LIGHTNING-SAFE ARE YOUR BURIED
 ON DISCONNECTING A LONG LINE.                CLASSICAL
RGE ANALYSIS OF OVERHEAD LINE CABLE SYSTEM WITH SKIN EF
OF OVERHEAD TRANSMISSION LINES.
       SHIELDING 400 KV LINES AGAINST DIRECT LIGHTNING
LTAGE ON DISCONNECTING A LONG LINE.                CLAS
```

Figure 9a. Permuted Title Index Showing Typographical Errors.

```
003  ELEC ENG 49(4): 33-4 (APR 1972)                    SHIE
005  ELEC ENG JAP 90(4): (JUL-AUG 1971)                 SWIT
015  ELEC ENG JAP 91(4): 201-8 (JUL-AUG 1971)           SPAR
001  ELEC ENG JAP 91(4): 69-78 (JUL-AUG 1971)           DIGI
004  ELEC INDIA 11(6): 31-5 (JUN 1971)                  TREN
008  ELEC LIGHT POWER 49(8): 42-4 (MAY 1971)            HOW
002  ELEC REV 191(1): 16 (JUL 7, 1972)                  PIPE
006  ELEC TECHNOL 4: (1970)                             CLAS
013  ELECT COMMUN 40(3): 381-4 (1965)                   ESTI
023  ELECT COMMUN 40(3): 381-4 (1965)                   ESTI
022  ELECT ENGNG JAPAN 84(7): 47-56 (JUL 1964)          SURG
026  ELECT ENGNG JAPAN 86(2): 39-46 (FEB 1966)          SURG
025  ELECT ENGNG JAPAN 86(7): 62-71 (JUL 1966)          SURG
018  ELECT ENGR 39(5): 37-43 (AUG 10, 1962)             LIGH
020  ELECT TIMES VOL 137: 409-10 (MAR 17); VOL          SURG
010  ELECTR WORLD 179(9): 42-5 (MAY 1, 1973)            SURG
024  ELEKT STANTSII (12): 85-6 (DEC 1966) (IN R         EXPE
012  ELEKTR STANTSII (11): 67-9 (NOV 1973) (IN          AN O
014  ELEKTRICHESTUO (4): 80-3 (1969) (IN RUSSIA         THE
028  ELEKTRICHESTUO (4): 80-3 (1969) (IN RUSSIA         THE
009  ELEKTRICHESTVO (1): 28-35 (JAN 1964) (IN R         LIGH
011  ELEKTRICHESTVO (3): 67-70 (MAR 1973) (IN R         THE
019  ELEKTRIE 14(5): 163-5 (MAY 1960) (IN GERMA         REPR
017  ELEKTRIE 17(1): 21-4 (JAN 1963) (IN GERMAN         MEAS
027  ELEKTRIE 23(10): 431-3 (OCT 1969) (IN GERM         TEST
```

Figure 9b. Journal Citation Errors and Inconsistencies.

Need for Standards

We now turn to a discussion of those areas of machine searching that need improvement. One possible solution to the problem of differences in data bases and suppliers is uniform standards. Standards allow the searcher to plan searches intelligently. They remove the haphazard approach of finding response "slow," data cells "down" (so that full bibliographic data cannot be printed), or files not updated because the supplier has not "gotten around to it yet." A more fundamental standard would be to announce that the goal for a system is to be up and available, say, 99.5 percent during scheduled hours.

More rigid standards are also needed in the data bases themselves. Figure 10 is a dictionary display from CA CONDENSATES using the DIALOG system. We are expanding around the term *cross*. Note that terms E1, E2, E36, E40-E43, E45, and E9-E12 are misspellings of proper variants of *cross-link*. It is reasonable that such errors happen, but is it reasonable that they remain uncorrected? The problem is one of divided responsibility—there is little motivation for the search vendors to fix up someone else's files. If Chemical Abstracts Service were to correct the data base, the corrected version would have to be reloaded by the search vendors at considerable cost. These spelling mistakes are potentially serious, since most on-line searching systems are based solely on string matching techniques. We would also make a plea for standards for treatment of authors' initials, uniform abbreviations for periodicals, etc. The number of variations in these from one data base to another is astonishing.

The Information Utility

The 1974 conference of the American Society for Information Science had as its theme "Information Utilities." This approach views bibliographic data base publishers and search system vendors as being part of a chain which in many respects resembles a public utility. It is therefore reasonable to state that these organizations should be accountable to their users. It should also be noted that many data base publishers enjoy pre-eminences in certain technical areas. It is unlikely that any group could challenge Chemical Abstracts Service or INSPEC and produce a competitive product. Indeed, such duplication would be wasteful and almost certainly counterproductive. It is better to have one excellent source of chemical bibliography than two good sources. Because of this pre-eminence, however, there should be strong mechanisms to ensure the maintenance of high and compatible standards throughout the information industry. Search vendors should be persuaded to establish, publish, and follow standards which minimize the user's problems and ensure maximum compatibility, service to service and data base to data base.

Ref	Index-term	Type	Items	RT
→ E1	CROSLINKED		1	0
→→ E2	CROSLINKING		1	0
E3	CROSS		4613	0
E4	CROSSARM		1	0
E35	CROSSINGS		12	0
→→ E36	CROSSINKING		1	0
E37	CROSSITE		3	0
E38	CROSSLAID		1	0
E39	CROSSLAND		1	0
→ E40	CROSSLIINK		1	0
→ E41	CROSSLIINKING		2	0
→ E42	CROSSLINBING		1	0
→ E43	CROSSLING		1	0
E44	CROSSLINK		263	0
→ E45	CROSSLINKAABLE		1	0
E46	CROSSLINKABILITY		2	0
E47	CROSSLINKABLE		119	0
E48	CROSSLINKAGE		33	0
E49	CROSSLINKAGES		8	0
E50	CROSSLINKED		1049	0
E51	CROSSLINKER		271	0
E6	CROSSLINKERS		8	0
E7	CROSSLINKING		3956	0
E8	CROSSLINKINGS		4	0
→ E9	CROSSLINKINNG		1	0
→ E10	CROSSLINKKING		1	0
→ E11	CROSSLINKLING		1	0
→ E12	CROSSLINKNG		1	0
E13	CROSSLINKS		97	0
E14	CROSSMIXING		1	0
E15	CROSSOPTERYGIAN		1	0
E16	CROSSOVER		54	0
E17	CROSSOVERS		4	~

Figure 10. DIALOG EXPAND Listing Showing Misspellings in CA CONDENSATES Data Base

We can now draw some general conclusions. Certainly, the computer has had a major impact on the business of information retrieval, and it has made its mark at Bell Laboratories. Machine searching has made many advances in the past two years. If used judiciously and with full awareness of the pitfalls it offers many advantages. It saves time and money, it is comprehensive, and it can be fast.

However, machine searching is not necessarily the best method in all

cases. Many searches still have to be done manually, such as those going back in time beyond the scope of the available data bases, and those not well defined, requiring browsing of a data base or index. (The eye can skim a page much faster than a 30cps terminal can print. Serial presentation can never surpass the random retrieval of the eye.)

On-line systems generally give better results than batch systems. They offer the searcher the ability to learn from intermediate results as the search proceeds. They also offer the opportunity to combine terms—in a Boolean sense—in ways which are impossible using printed indexes. However, machine searching systems suffer from errors and inconsistencies in the data bases, as well as from a lack of standards. As vendors become more aware of these shortcomings, we are confident that improvements will occur, and we look forward to an increased exploitation of what is already an indispensable tool of a literature searching service.

REFERENCES

1. Hawkins, Donald T. "Bibliographic Data Base Usage in a Large Technical Community," *Journal of the American Society for Information Science* 25:105-08, March-April 1974.

2. Cuadra, Carlos A. "SDC Experiences with Large Data Bases," *Journal of Chemical Information and Computer Sciences* 15:48-51, Feb. 1975.

3. Elman, Stanley A. "Cost Comparison of Manual and On-Line Computerized Literature Searching," *Special Libraries* 66:12-18, Jan. 1975.

4. Hawkins, Donald T. "A Bibliography on the Physical and Chemical Properties of Water, 1969-1974," *Journal of Solution Chemistry* 4:623-743, Aug. 1975.

5. Barber, A. Stephanie, *et al.* "On-line Information Retrieval as a Scientist's Tool," *Information Storage and Retrieval* 9:429-40, Aug. 1973.

6. Lowry, W. Kenneth. "Use of Computers in Information Systems," *Science* 175:841-46, Feb. 25, 1972.

7. Schipma, Peter B. "Searching Costs [letter]," *Special Libraries* 65:6A, Aug. 1974.

8. Weiner, Norbert. *The Human Use of Human Beings; Cybernetics and Society*. Rev. ed. Garden City, N. Y., Doubleday, 1954.

9. Deutsch, L. Peter, and Lampson, Butler W. "An On-line Editor," *Communications of the ACM* 10:793-99, Dec. 1967.